**ALY 6020 PREDICTIVE ANALYTICS**

**CRN -22804**

**“TOPIC MODELING HOMEWORK”**

**RISHABH SHARMA (sharma.ris@northeaster.edu)**

**CPS-2021 WINTER QUARTER**



**ASSIGNMENT DATE- 8 APRIL 2020**

**INSTRUCTOR- Prof.** **ABHIJIT SANYAL**

**INTRODUCTION**

Our dataset consists of 3 books namely Twenty Thousand Leagues under the Sea (164), The War of the Worlds (36), and Wuthering Heights (768) which we will extract using the command “**gutenberg\_download**”. We will perform TDA, LDA methods to achieve our results.

Text

Description automatically generated

**ANALYSIS**

For finding the theoretical subjects that happen in an assortment of records we utilized the point model which is a kind of measurable model. By and large, point displaying is utilized as text-digging apparatus for disclosure of covered up semantic constructions in a content body, and subjects delivered by this demonstrating are bunches of comparable words.

Presently, the **Latent Dirichlet Allocation (LDA)** is a famous technique that sudden spikes in demand for the head of treating each record are a combination of themes and each subject is a combination of words. In this displaying we could accept on a 3-point model that a report contains 70% about theme 1, 30% about subject 2, and 0% about subject 3, and a subject is a probabilistic conveyance over the different word.

We here come to three archives that are **Twenty Thousand Leagues under the Sea, The War of the Worlds** and **Wuthering Heights** from Gutenberg with report term lattice (DTM) to examine our theme displaying. We executed a subject model with **K = 3** i.e., 3 books for every theme to recognize three unique books by applying Latent Dirichlet Allocation (LDA). In the wake of separating the three books together, we partition them into records addressing one section each and tokenize the dataset of every one of the three books together, and finally eliminating the stop words. We have made archive term lattice and applied LDA to make 3-point displaying and we found that the record term grid made from word checks information outline has 95% sparsity which implies our dataset doesn't have a scope of estimations of at least one measurement.Text

Description automatically generated

After applying LDA with 3-Topic demonstrating we tracked down the main 5 terms utilized in the general dataset and found that Martians, individuals, dark, time, and street shows up most in the record or every subject. Furthermore, after plotting or getting a representation of these best 5 probabilities for every point we discovered words for every theme.**Chart, bar chart

Description automatically generated**

Point 1 has words Martians, individuals, dark, time, and street and on the off chance that I needed to figure which book this theme had a place with, I would say the War of the Worlds or it could even be associated with Wuthering Heights. Theme 2 has terms identified with the ocean like a chief, ocean, Nautilus, Nemo, and ned so it is reasonable to think it has a place with Twenty Thousand Leagues under the Sea. Point 3 has two names Catherine, and Linton, and the other three terms are I'm, I'll, and Heathcliff. Very little to pass by, except if one knows the substance of the books.

Presently by utilizing the Gamma grid, we discovered how every theme is connected with each report and we can say these qualities are an expected extent of words from each archive that are created from that point. As shown below, the model gauges that each word in the Twenty Thousand Leagues under the Sea\_26 record has just a 0% likelihood of coming from subject 1 (Twenty Thousand Leagues under the Sea). Table

Description automatically generated

We have seen below that every one of the sections from Twenty Thousand Leagues under the Sea, The War of the Worlds, and Wuthering Heights were interestingly distinguished as a solitary subject each.

**Chart

Description automatically generated**

As anticipated before, for points one and two, we can see here that they are bound to come from the books Twenty Thousand Leagues under the Sea, The War of the Worlds, individually. For subject three, it would likely be found in Wuthering Heights even though there are some covering focuses with point one. In general, the calculation did well in anticipating the subjects to the comparing books.

We can see from the image below all part has a place with the theme 1 i.e., the conflict of the universes and according to my previous theories, I referenced point one as conceivably has a place with the conflict of the universes. **Table

Description automatically generated**

Presently, the mix of the genuine book (title) and the book allocated to it (agreement) is helpful for additional investigation. We can see structure beneath the perception of disarray lattice, we see that every one of the words for Twenty Thousand Leagues under the Sea, The War of the Worlds, and Wuthering Heights were accurately relegated and don't have quite a few misassigned words.Chart, waterfall chart

Description automatically generated

Presently, when our theme model isn't performing at its best situation then I would prescribe the accompanying interesting points for improving the interpretability of the subject demonstrating.

1. The absolute first thing is to recognize phrases through n-grams and channel thing type structures. It like distinguishing the expressions for the subject model to remember them effectively like expressions containing 2 words or 3 words e.g., online media or Tic Tac Toe, and this will be conceivable utilizing Pointwise Mutual Information (PMI) score which estimates co-happening of words than if they are free.

2. Filtering leftover words for things like in the sentence ' Taj Mahal is excellent', we realize the sentence is discussing Taj Mahal which means words in the sentence giving more clarity of mind and clarification about the point (Taj Mahal) itself. Henceforth, separating for thing extricates words that are more interpretable for the theme model.

3. We can likewise streamline decisions for several points through cognizance measures like in LDA demonstrating it requires indicating the number of themes. We can change the number of points to advance the actions like prescient probability, perplexity, and lucidness.

4. Adjusting LDA boundaries like expanding passes, and cooperation’s, lump size to the degree of memory dealing with.

5. And the latter is TF-IDF which means **'Term Frequency-Inverse Document Frequency**' and it is a procedure to evaluate a word in reports. In this procedure, we utilize the calculation of weight to each word which implies the significance of the word in the record. Like 'New York is an extraordinary city and its simple for every one of us to comprehend the significance of sentence as we probably are aware the semantics of each word yet for PCs which comprehended number as it were. This procedure makes things conceivable by vectorizing the entirety of the content with the goal that the PC can comprehend the content better and mastermind in the positioning request by doing grouping or discovering the important records, etc.

**REFERENCES**

1. Code provided in the class module for reference.
2. Scott, W. (2019, May 21). TF-IDF from scratch in python on real world dataset. Medium. https://towardsdatascience.com/tf-idf-for-document-ranking-from-scratch-in-python-on-real-world-dataset-796d339a4089
3. Ruchirawat, N. (2020, November 1). 6 Tips for Interpretable Topic Models - Towards Data Science. Medium. https://towardsdatascience.com/6-tips-to-optimize-an-nlp-topic-model-for-interpretability-20742f3047e2